

TEXT TO SPEECH SYSTEM AND METHOD HAVING INTERACTIVE
SPELLING CAPABILITIES

5

TECHNICAL FIELD OF THE INVENTION

10

This invention relates generally to text-to-speech technology and more particularly, to a text-to-speech system and method having interactive spelling capabilities.

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191

BACKGROUND OF THE INVENTION

Systems using adaptive speech recognition (ASR) and text-to-speech (TTS) technology allow users to interact with computer systems using spoken commands. Such systems greatly increase the flexibility and accessibility of computer systems, allowing users to reach computer systems anywhere they can access a telephone. As more people come to rely on these systems, users have demanded even greater functionality in ASR and TTS technology.

5

SUMMARY OF THE INVENTION

In accordance with the present invention, techniques for relating audio files with text files, as well as using the relation between the audio and text files for spelling out words, are provided. In accordance with the present invention, the disadvantages and problems associated with identifying words in text files corresponding to words in audio files have been substantially reduced or eliminated.

In accordance with one embodiment of the present invention, a method for audibly spelling a word in an audio file includes playing an audio file to user, and receiving from the user a command to spell a word in the audio file. The method also includes identifying a textual word corresponding to the word to be spelled in a text file, and audibly spelling the textual word.

In accordance with another embodiment of the present invention, an interactive voice response system (IVR) includes an interface and a processor. The interface plays an audio file to a user, and receives a command to spell a word in the audio file from the user. The processor identifies the word to be spelled, identifies a textual word in a text file corresponding to that word, and audibly spells the word.

In accordance with another embodiment of the present invention, a method for relating an audio file to a text file includes retrieving a text file that includes written words and generating an audio file from the text file. The method further includes storing information relating audible words in the audio file to corresponding textual words in the text file.

In accordance with yet another embodiment of the present invention, a text-to-speech (TTS) system includes a

memory and a TTS module. The memory stores the text file made of textual words and also stores an audio file made of audible words. The TTS module uses the text file to generate an audible word corresponding to each textual word in the text file and to store these audible words in an audio file. The TTS module then stores a file map in the memory that includes a first location for each audible word in the audio file, and a second location for each corresponding textual word in the text file. Accordingly, the file map synchronizes the audible words in the audio file with the textual words in the text file.

Technical advantages of the present invention include the ability to locate a textual word in a text file corresponding to an audible word. Accordingly, a system may perform processing on a textual word corresponding to an audible word. For example, a system could receive a command from a user to spell an audible word played to the user. The system could then spell the word by locating the corresponding textual word and pronouncing the individual characters of that word. Such a system is useful when the user cannot determine the spelling of the word by hearing the word, such as when the user hears a name derived from a language foreign to the user.

Another technical advantage of certain embodiments of the present invention is allowing the user with voice access to a system to distinguish words that sound the same but are spelled differently. Accordingly, the system can audibly spell the word to allow the user to determine which spelling is correct. For example, if a user of a voice-operated stock trading system wishes to verify that the user is purchasing stock in Cisco (a prominent technology supplier) rather than SYSCO (a prominent food supplier), the user

5

BRIEF DESCRIPTION OF THE DRAWINGS

For a more complete understanding of the present invention and its advantages, reference is now made to the following description taken in conjunction with the accompanying drawings, in which:

FIGURE 1 illustrates a communication system that enables interactive spelling during playback of an audio file;

FIGURE 2 illustrates a file map generated by the system of FIGURE 1, which associates text of a text file with audio of an audio file; and

FIGURE 3 is a flowchart illustrating a method for interactive spelling of words during playback of an audio file, in accordance with a particular embodiment of the present invention.

DETAILED DESCRIPTION THE INVENTION

FIGURE 1 illustrates a particular embodiment of a communication system 100 that creates audio files from text files and stores information relating audible words in the audio files to corresponding textual words in the text files. System 100 can therefore perform any suitable processing task that involves identifying a textual word corresponding to an audible word. For example, system 100 may play an audio file to a user of communication device 102, receive a command to spell a word in the audio file from the user, and spell the word in response to the command using the corresponding textual word.

System 100 is a text-to-speech (TTS) system that generates audio files 122 from text files 120 stored in a memory 114, allowing system 100 to read text files 120 aloud to a user with voice access to system 100. The term "text-to-speech" refers to any of a number of methods of converting electronic text into an audio form. TTS technology allows a user to access text information in system 100 without requiring a visual display. For example, the user could contact system 100 by telephone and instruct system 100 to read email to the user. System 100 may also use adaptive speech recognition (ASR) to receive voice commands from the user and to take appropriate action in response to the commands. ASR technology allows system 100 to receive speech from the user and to parse the speech into recognizable grammar, words, or vocabulary. For example, the user could say the word "spell," and system 100 could translate the word into a command recognizable by system 100.

Text files 120 used in system 100 may include letters, numbers, symbols, nonprinting characters, and any other

5 suitable text information, some or all of which may be formed into textual words 126. Textual words 126 include any electronic representation of one or more recognizable characters that can be pronounced in a language appropriate to the characters. One example of a textual word 126 is a string of alphabetic ASCII characters.

10 Audio files 122 include audible words 128 that are audible pronunciations of textual words 126 formed from synthesized or recorded sounds. Audible words 128 may be separated by pauses or periods of silence to indicate the end of a sentence, the presence of a comma or dash, or any other suitable format or punctuation. The exact pronunciation and intonation of the words may be varied based on punctuation marks or other indications in text file 120. For example, if textual word 126 preceded a question mark, the corresponding audible word 128 may be pronounced with a higher pitch, a vocal inflection that is commonly used in conversation to indicate that a query is being posed. Audible words 128 may be stored in any appropriate electronic format, including WAV files, MP3 files, or any other suitable form of electronic storage.

20 In order to allow system 100 to identify a particular textual word 126 corresponding to an audible word 128 and vice versa, system 100 stores information that relates the locations of audible words 128 within audio file 122 to the locations of textual words 126 within text file 120. System 100 locates audible words 128 and textual words 126 using any suitable method for locating information within a file appropriate to the format of the file. For example, textual word 126 may be located in a ASCII text file by numbering the characters in the text file and recording the number of the first and last characters of the textual word. Locations

of audible words 128 and textual words 126 may be stored in a file map 124 that relates each audible word 128 to a corresponding textual word 126.

5 In an alternative embodiment, the positions of audible words 128 are marked within audio file 128 itself using voice tags 130. Voice tags 130 are electronic markers within audio file 128 that indicate the position of words within audio file 128. In a particular embodiment, system 100 removes voice tags 130 during playback and keeps track of the number of words played by counting the number of voice tags 130 removed.

10 System 100 may use the information in file map 124 to locate textual words 126 corresponding to audible word 128 and to perform suitable processing on textual word 126. In a particular embodiment, system 100 may spell audible word 122 by identifying the corresponding textual word 126 and pronouncing the individual characters of the corresponding textual word 126. Audible spelling is a useful feature when the user of system 100 has only voice access to system 100 and thus cannot distinguish words that sound alike but are spelled differently, such as "night" and "knight," or in any other circumstances where the user has difficulty determining the spelling of a word. For example, if the name of a person or a street is derived from a language 15 foreign to the user, the user might not be able to determine the spelling of the name from its pronunciation alone. Consequently, such a user would find it advantageous to hear the spelling of the name as well.

20 The components of system 100 are described in greater detail below. As depicted in FIGURE 1, system 100 includes an interface 108 that communicates with a communication device 102, an adaptive speech recognition (ASR) module 110, 25 30

a processor 112, a memory 114 and a text-to-speech (TTS) module 116. Although system 100 is illustrated with particular components, it is understood that the functions of various components of system 100 may be performed by single components or distributed amongst several components and that components of system 100 may be omitted, replaced or added without substantially changing the overall operation of system 100.

Communication device 102 allows a user to communicate with system 100 using speech. Device 102 converts the speech received from the user into an electronic form and communicates that information to system 100. Device 102 includes a microphone 104 that receives speech from the user and a speaker 106 that plays back audio information from system 100 to the user. Device 102 may also include a keypad 105 or other suitable input device for receiving information from the user.

Interface 108 receives information from communication device 102 and communicates information from system 100 to device 102. Interface 108 includes any port or connection, real or virtual, that allows system 100 to exchange information with communication device 102. Interface 108 may receive and transmit information using wires, or may be a wireless interface that decodes wireless signals into information for system 100.

ASR module 110 interprets speech received from the user into text and commands. ASR module 110 may be implemented as hardware, software, or any combination of the two. ASR module 110 communicates with other components of system 100; for example, ASR module 110 may communicate commands to processor 112.

Processor 112 represents a microprocessor, microcontroller, digital signal processor (DSP), or any other hardware and/or software configured to process information. Processor 112 manages components of system 100. Processor 112 can also execute commands received from the user of system 100, such as commands parsed from the user's speech by ASR module 110 or entered using keypad 105.

Memory 114 represents any volatile or non-volatile memory, including magnetic media, optical media, read-only memory (ROM), random access memory (RAM), removable media, CD-ROMs, DVD-ROMs, or any other suitable local or remote memory component. Memory 114 stores text files 120, audio files 122, and file maps 124 relating text files 120 to audio files 122. Memory 114 may be accessed by one or more components of system 100 to store information. For example, ASR module 110 may convert speech from the user into electronic form and store the speech as a text file 120 in memory 114.

TTS module 116 represents any hardware and/or software that generates audio files 122 from text files 120 and stores information relating audible words 128 in audio file 122 to textual words 126 in text file 120. TTS module 116 performs any suitable processing to identify textual words 126, to generate audible words 128, and to store audible words 128 in any suitable electronic form. TTS module 116 may also play back audio file 128 by converting audio file 128 into media stream packets for any suitable audio output device. Media stream packets include any packet, cell, frame, segment or portion of data suitable for storing and communicating audio information in electronic form. For example, TTS module 116 could convert audio file 128 into media stream packets that are transmitted to device 102 via

interface 108. To audibly spell textual words 126, TTS module 116 identifies the characters in a textual word 126, determines a pronunciation for each of the characters, and uses the pronunciation of the characters to generate electronic information, control signals or any other suitable form of audio information corresponding to the pronunciation of the characters. In another embodiment, TTS module 166 retrieves an electronically stored sound corresponding to the character, such as "A," "B," or "C," and plays the sound back to the user.

In one example of system 100 in operation, system 100 retrieves a text file 120 from memory 114. Text file 120 may be created by any source including local components of system 100 as well as remote systems. In a particular embodiment, text file 120 is generated from email, content of a database, content of web pages, or other suitably stored information. Alternatively, text file 120 may be generated from the speech of a user of system 100 by ASR module 110. System 100 uses TTS module 116 to generate audible words 128 corresponding to the textual words 126 in text file 120 and to store audible words 128 in an audio file 122. TTS module 116 also generates file map 124 that includes locations for each audible word 128 within audio file 122 and corresponding locations of textual words 126 within text file 120. Accordingly, file map 124 provides a mechanism for synchronizing audio file 122 with text file 120.

After audio file 122 is created, system 100 converts audio file 122 and plays audio file 122 to user using speaker 106. During playback, user might hear a word that user wants to have spelled by system 100. In the particular embodiment shown, for example, the user may wish to

determine whether "Bryan" in "Bryan Street" is spelled "Bryan" or "Brian." Or the user may want to clarify whether system 100 played back "Go to the light" rather than "Go to the right."

5 To have system 100 spell a word, the user utters a command to spell the word into microphone 104. Because the command is received during playback, it is referred to as a "barge-in command." ASR module 110 interprets the command and instructs processor 112 to spell the word in response to the command. System 100 then halts playback of audio file 122, and determines the last audible word 128 that was played from audio file 122. System 100 determines a textual word 126 in text file 120 corresponding to the audible word 128 last played from audio file 122 using file map 124. 10 Once the proper textual word 128 has been located, TTS module 126 can spell out the textual word 128 by pronouncing the individual characters of the textual word 128. 15

Once the word has been spelled, system 100 may receive follow-up commands from the user. If a word has been misspelled, the user may utter a command to correct the spelling of the word, and system 100 then receives the correct spelling from the user and modifies the spelling in text file 120. If the user prefers to have another word spelled, the user utters a command such as "previous", "next", "entire paragraph," or any other appropriate direction for system 100. System 100 then selects a new word in text file 120, and spells the word. Alternatively, system 100 may play a prompt to user which includes the word, such as "Is <word> the word you want to spell?" 20 System 100 then receives feedback from the user and either spells the word or selects a new word. Once the user has received the spelling for each of the words that the user 25 30

wishes to have spelled, the user can utter a "resume" command that instructs system 100 to resume playing audio file 122 where it left off previously.

5 In another example of system 100 in operation, system 100 converts each textual word 126 of text file 120 into a separate audio file 122. In this alternative mode of operation, system 100 stores file map 124 that identifies the audio file 122 that corresponds to each textual word 126. Instead of playing a single audio file 122 to user, system 100 plays a sequence of audio files 122 corresponding to the order of words in text file 120. When the user requests the spelling for a word, system 100 determines the last audio file 122 that was played before the command was received. System 100 then determines textual word 126 corresponding to the last audio file 122 played using file map 124, and plays the spelling of that textual word 126 to the user.

10
15
20 In still another example of system 100 in operation, system 100 generates media stream packets from text file 120 and plays them back continuously, rather than storing all of the audio information in audio file 122 before playing the audio information back. Conceptually, this is analogous to system 100 reading text file 120 aloud in real time. In this specific embodiment, audio information is stored only briefly before it is played back, so audio file 122 functions as a virtual file, a transitory form of electronic storage for media stream packets. Since audio information is stored only transitorily, there is no need to store information relating audio file 122 to text file 120. Instead, system 100 need only determine the last text word 126 from which an audible word 128 was generated. For example, system 100 could use a counter to keep track of the

number of text words 126 in text file 122 that have been "read" to the user in the form of audible words 128. In other embodiments, system 100 may store more complicated information, such as a file map 124 relating each textual word 126 to a time at which the corresponding audible word 128 was played.

System 100 and its modes of operation are examples of an interactive spelling system, but numerous variations and modifications will be apparent to one skilled in the art. For example, instead of speaking commands to system 100, the user could enter them directly using keypad 105 or any other suitable input device. Other modifications include more complicated text processing, such as allowing the user to insert new text or perform more complicated editing functions. Because system 100 may be modified in a number of ways without substantially affecting the overall functions of system 100, the description of system 100 should be viewed as illustrative rather than restrictive.

FIGURE 2 illustrates a sample file map 200 for text file 120 and audio file 122 depicted in FIGURE 1. File map 200 includes for each word a word number 202 identifying the word 126, a first character 204, a last character 206, a beginning audio time 208, and an ending audio time 210. Word number 202 corresponds to the ordering of the textual words 126 within text file 120, and identifies each textual word 126 in text file 120. For example, the fifth word in the sample passage is "house," so its word number 202 is five. First character 204 indicates the character number within the text file 120 at which each word starts, and last character 206 is a character at which it ends. Similarly, beginning audio time 208 and ending time 210 indicate the

time at which corresponding audible words 128 within audio file 122 will be played during playback.

File map 200 provides information that would not be available from text file 120 alone. For example, the word "Street" is at the end of a sentence, so it is followed by a period and two spaces before the next word "Go," as indicated by the difference between last character 206 of "Street" and first character 204 of "Go." Because of the period, the separation between the corresponding audible words 128 is somewhat longer to reflect the pause introduced by the punctuation, as indicated by the pause between the ending time 210 of "Street" and the beginning time 208 of "Go." It can be difficult, however, to ascertain the exact length of a particular pause solely from the character separation. Similarly, the duration of particular audible words 128 may be only loosely correlated with the number of characters in the corresponding textual words 126. For example, "Street" has more characters than "Bryan," but "Bryan" takes longer to pronounce because it has an additional syllable. These examples illustrate some of the difficulties associated with attempting to calculate the correspondence between audio files 122 and text files 120 using only the information in a text file 120. But file map 200 provides exact positions for textual words 126 and audible words 128, thus obviating the need for error-prone calculations to locate audible words 128 within audio file 122.

The information in file map 200 can be suitably modified by one skilled in the art. For example, locations in audio file 122 could be labeled by byte numbers or other identifiers that specify a memory location within audio file 122. The byte numbers could then be used to locate

particular audible words 128 within audio file 122. In another modified form, the structure of file map 200 may be different from the depicted table. For example, system 100 may use voice tags 130 or other labels within audio file 122 itself that indicate the locations of audible words 128. File map 200 should not be construed to exclude such modifications; rather, it should be taken as an illustrative example of such possibilities.

FIGURE 3 is a flowchart 300 illustrating a method for interactive spelling using system 100. System 100 retrieves text file 120 at step 302. TTS module 116 generates audio file 122 from text file 120 at step 304, and also stores file map 124 relating audio file 122 to text file 120 at step 305. At step 306, system 100 plays audio file 122 to the user of system 100.

During playback, system 100 monitors for commands from the user at step 308. The system determines if a command has been received at step 310. If no command has been received, system 100 determines whether audio file 122 has been played completely at step 311. If audio file has not been completely played, system 100 continues to play audio file 122 and monitor for commands during playback at step 308.

If a spell command is received, system 100 halts playback of audio file 122 at step 312. System 100 then determines the last audible word 128 that was played from audio file 122 at step 314. Using file map 124, system 100 identifies textual word 126 corresponding to the last audible word 128 played at step 316. System 100 plays the spelling of textual word 126 to user at step 318. System 100 then monitors for another command at step 320.

At step 322, system 100 determines whether it has received a command to replay the spelling. If system 100 receives a replay command, system 100 replays the spelling at step 318, and then continues to monitor for new commands at step 320. If system 100 does not receive a replay command, system 100 determines whether it has received a command to resume playback of audio file 122 at step 324. If a command to resume has been received, then at step 326, system 100 resumes playing audio file 122 where it left off. During playback, system 100 continues to monitor for new commands at step 308.

If neither a replay nor a resume command is received, system 100 determines if it received an edit command at step 328. If system 100 receives an edit command, system 100 receives a corrected spelling from the user at step 330. At step 332, system 100 verifies the corrected spelling, such as by playing the spelling back to the user, and system 100 then stores the corrected spelling in text file 120 at step 334. After the corrected spelling is stored, system 100 updates file map 124 to reflect the correction to text file 120 if necessary. For example, file map 124 could be updated if the number of characters in the corrected textual word 126 is different from the original textual word 126. Once the corrected spelling is stored and file map 124 is updated, system 100 monitors for the next command at step 320.

If none of the above commands are received, system 100 interactively selects a new word to spell at step 334. For example, system 100 could prompt the user to select a new word and respond to commands such as "previous" or "next." Once the new word is selected, system 100 plays the spelling

of the new word at step 318, and continues monitoring for commands at step 320.

5 Numerous variations and modifications of the method presented will be apparent to one skilled in the art. For example, additional commands could be added, and particular steps in the method could be omitted, added, rearranged, or replaced without substantially affecting the overall operation of the method. Consequently, flowchart 300 should be viewed as an example of a method having interactive spelling capabilities rather than as an exclusive definition of such a method.

10
15
20
25
30
FIGURE 4 is a flowchart 400 illustrating one embodiment of a method for relating text files 120 to audio files 122. System 100 retrieves text file 120 at step 402. Text file 120 may be stored locally or remotely and may be generated from email, retrieved from a database or a web page, input by a user, or produced by any other suitable method. System 100 generates audio file 122 from text file 120 at step 404. At step 406, system 100 stores information relating audible words 128 in audio file 122 to textual words 126 in text file 120. System 100 determines whether another text file 120 needs to be converted at step 408. If so, system 100 can repeat the method from step 402 for the next text file 120.

25 Again, the method is subject to numerous modifications by a person skilled in the art. Various steps of the method may be added, omitted, rearranged, replaced, and performed concurrently and/or continuously without substantially affecting the overall operation of the method.
30 Consequently, flowchart 400 should be viewed as one of many possible examples of a method for relating textual words 126 in text file 120 to audible words 128 in audio file 122.

Although the present invention has been described with several embodiments, a myriad of changes, variations, alternations, transformations, and modifications may be suggested to one skilled in the art, and it is intended that the present invention encompass such changes, variations, alternations, transformations and modifications as fall within the scope of the appended claims.

5

10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995